Reassessing the Goals of Grammatical Error Correction: Fluency Instead of Grammaticality

Keisuke Sakaguchi[1], **Courtney Napoles**[1], **Matt Post**[2], and **Joel Tetreault**[3]
[1]Center for Language and Speech Processing, Johns Hopkins University
[2]Human Language Technology Center of Excellence, Johns Hopkins University
[3]Yahoo
{keisuke,napoles,post}@cs.jhu.edu, tetreaul@yahoo-inc.com

Abstract

The field of grammatical error correction (GEC) has grown substantially in recent years, with research directed at both evaluation metrics and improved system performance against those metrics. One unvisited assumption, however, is the reliance of GEC evaluation on *error-coded* corpora, which contain specific labeled corrections. We examine current practices and show that GEC's reliance on such corpora unnaturally constrains annotation and automatic evaluation, resulting in (a) sentences that do not sound acceptable to native speakers and (b) system rankings that do not correlate with human judgments. In light of this, we propose an alternate approach that jettisons costly error coding in favor of unannotated, whole-sentence rewrites. We compare the performance of existing metrics over different gold-standard annotations, and show that automatic evaluation with our new annotation scheme has very strong correlation with expert rankings ($\rho = 0.82$). As a result, we advocate for a fundamental and necessary shift in the goal of GEC, from correcting small, labeled error types, to producing text that has *native fluency*.

1 Introduction

What is the purpose of grammatical error correction (GEC)? One response is that GEC aims to help people become better writers by correcting grammatical mistakes in their writing. In the NLP community, the original scope of GEC was correcting targeted error types with the goal of providing feedback to non-native writers (Chodorow and Leacock, 2000; Dale and Kilgarriff, 2011; Leacock et al., 2014). As systems improved and more advanced methods were applied to the task, the definition evolved to *whole-sentence correction*, or correcting all errors of every error type (Ng et al., 2014). With this pivot, we urge the community to revisit the original question.

It is often the case that writing exhibits problems that are difficult to ascribe to specific grammatical categories. Consider the following example:

Original: From this scope , social media has shorten our distance .
Corrected: From this scope , social media has shortened our distance .

If the goal is to correct verb errors, the grammatical mistake in the original sentence has been addressed and we can move on. However, when we aim to correct the sentence as a whole, a more vexing problem remains. The more prominent error has to do with how unnaturally this sentence reads. The meanings of words and phrases like *scope* and the corrected *shortened our distance* are clear, but this is not how a native English speaker would use them. A more fluent version of this sentence would be the following:

Fluent: From this perspective , social media has shortened the distance between us .

This issue argues for a broader definition of grammaticality that we will term *native-language fluency*, or simply *fluency*. One can argue that traditional understanding of grammar and grammar correction encompasses the idea of native-language fluency. However, the metrics commonly used in evaluating GEC undermine these arguments. The performance of GEC systems is typically evaluated us-

Transactions of the Association for Computational Linguistics, vol. 4, pp. 169–182, 2016. Action Editor: Chris Quirk.
Submission batch: 12/2015; Revision batch: 4/2015; Published 5/2016.

ing metrics that compute corrections against *error-coded* corpora, which impose a taxonomy of types of grammatical errors. Assigning these codes can be difficult, as evidenced by the low agreement found between annotators of these corpora. It is also quite expensive. But most importantly, as we will show in this paper, annotating for explicit error codes places a downward pressure on annotators to find and fix concrete, easily-identifiable grammatical errors (such as *wrong verb tense*) in lieu of addressing the native fluency of the text.

A related problem is the presence of multiple evaluation metrics computed over error-annotated corpora. Recent work has shown that metrics like M^2 and I-measure, both of which require error-coded corpora, produce dramatically different results when used to score system output and produce a ranking of systems in conventional competitions (Felice and Briscoe, 2015).

In light of all of this, we suggest that the GEC task has overlooked a fundamental question: What are the best practices for corpus annotation and system evaluation? This work attempts to answer this question. We show that native speakers prefer text that exhibits fluent sentences over ones that have only minimal grammatical corrections. We explore different methods for corpus annotation (with and without error codes, written by experts and non-experts) and different evaluation metrics to determine which configuration of annotated corpus and metric has the strongest correlation with the human ranking. In so doing, we establish a reliable and replicable evaluation procedure to help further the advancement of GEC methods.[1] To date, this is the only work to undertake a comprehensive empirical study of annotation *and* evaluation. As we will show, the two areas are intimately related.

Fundamentally, this work reframes grammatical error correction as a fluency task. Our proposed evaluation framework produces system rankings with strong to very strong correlations with human judgments (Spearman's $\rho = 0.82$, Pearson's $r = 0.73$), using a variation of the GLEU metric (Napoles et al., 2015)[2] and two sets of "fluent" sentence rewrites as a gold standard, which are simpler and cheaper to collect than previous annotations.

2 Current issues in GEC

In this section, we will address issues of the GEC task, reviewing previous work with respect to error annotation and evaluation metrics.

2.1 Annotation methodologies

Existing corpora for GEC are annotated for errors using fine-grained coding schemes. To create error-coded corpora, trained annotators must identify spans of text containing an error, assign codes corresponding to the error type, and provide corrections to those spans for each error in the sentence.

One of the main issues with coded annotation schemes is the difficulty of defining the granularity of error types. These sets of error tags are not easily interchangeable between different corpora. Specifically, two major GEC corpora have different taxonomies: the Cambridge Learner Corpus (CLC) (Nicholls, 2003) has 80 tags, which generally represent the word class of the error and the type of error (such as replace preposition, unnecessary pronoun, or missing determiner). In contrast, the NUS Corpus of Learner English (NUCLE) (Dahlmeier et al., 2013) has only 27 error types. A direct conversion between them, if possible, would be very complex. Additionally, it is difficult for annotators to agree on error annotations, which complicates the annotation validity as a gold standard (Leacock et al., 2014). This is due to the nature of grammatical error correction, where there can be diverse correct edits for a sentence (Figure 1). In other words, there is no single gold-standard correction. The variety of error types and potential correct edits result in very low inter-annotator agreement (IAA), as reported in previous studies (Tetreault and Chodorow, 2008; Rozovskaya and Roth, 2010; Bryant and Ng, 2015).

This leads to a more fundamental question: why do we depend so much on fine-grained, low-consensus error-type annotations as a gold standard for evaluating GEC systems?

One answer is that error tags can be informative and useful to provide feedback to language learners, especially for specific closed-class error types

[1] All the scripts and new data we collected are available at `https://github.com/keisks/reassess-gec`.

[2] This metric should not be confused with the method of the same name presented in Mutton et al. (2007) for sentence-level fluency evaluation.

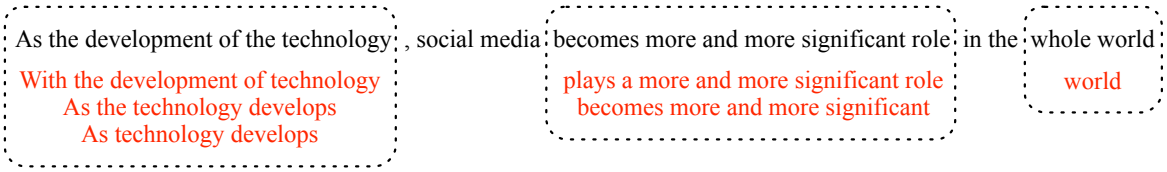

Figure 1: An ungrammatical sentence that can be corrected in different ways.

(such as determiners and prepositions). Indeed, the CLC, the first large-scale corpus of annotated grammatical errors, was coded specifically with the intent of gathering statistics about errors to inform the development of tools to help English language learners (Nicholls, 2003). Later GEC corpora adhered to the same error-coding template, if not the same error types (Rozovskaya and Roth, 2010; Yannakoudakis et al., 2011; Dahlmeier et al., 2013).

The first shared task in GEC aspired to the CLC's same objective: to develop tools for language learners (Dale and Kilgarriff, 2011). Subsequent shared tasks (Dale et al., 2012; Ng et al., 2013) followed suit, targeting specific error types. Error-coded corpora are effective training and evaluation data for targeted error correction, and statistical classifiers have been developed to handle errors involving closed-class words (Rozovskaya and Roth, 2014). However, the 2014 CoNLL shared task engendered a sea change in GEC: in this shared task, systems needed to correct all errors in a sentence, of all error types, including ones more stylistic in nature (Ng et al., 2014). The evaluation metrics and annotated data from the previous shared task were used; however we argue that they do not align with the use case of this reframed task. What is the use case of whole-sentence correction? It should not be to provide specific targeted feedback on error types, but rather to rewrite sentences as a proofreader would.

The community has already begun to view whole-sentence correction as a task, with the yet unstated goal of improving the overall *fluency* of sentences. Independent papers published *human* evaluations of the shared task system output (Napoles et al., 2015; Grundkiewicz et al., 2015), asking judges to rank systems based on their grammaticality. As GEC moves toward correcting an entire sentence instead of targeted error types, the myriad acceptable edits will result in much lower IAA, compromising evaluation metrics based on the precision and recall of coded errors. At this juncture, it is crucial that we examine whether error-coded corpora and evaluation are necessary for this new direction of GEC.

Finally, it would be remiss not to address the cost and time of corpus annotation. Tetreault and Chodorow (2008) noted that it would take 80 hours to correct 1,000 preposition errors by one trained annotator. Bryant and Ng (2015) reported that it took about three weeks (504 hours) to collect 7 independent annotations for 1,312 sentences, with all 28 CoNLL-2014 error types annotated. Clearly, constructing a corpus with fine-grained error annotations is a labor-intensive process. Due to the time and cost of annotation, the corpora currently used in the community are few and tend to be small, hampering robust evaluations as well as limiting the power of statistical models for generating corrections. If an effective method could be devised to decrease time or cost, larger corpora—and more of them—could be created. There has been some work exploring this, namely Tetreault and Chodorow (2008), which used a sampling approach that would only work for errors involving closed-class words. Pavlick et al. (2014) also describe preliminary work into designing an improved crowdsourcing interface to expedite data collection of coded errors.

Section 3 outlines our annotation approach, which is faster and cheaper than previous approaches because it does not make use of error coding.

2.2 Evaluation practices

Three evaluation metrics[3] have been proposed for GEC: MaxMatch (M^2) (Dahlmeier and Ng, 2012), I-measure (Felice and Briscoe, 2015), and GLEU (Napoles et al., 2015). The first two compare the changes made in the output to error-coded spans of the reference corrections. M^2 was the metric used

[3]Not including the metrics of the HOO shared tasks, which were precision, recall, and F-score.

for the 2013 and 2014 CoNLL GEC shared tasks (Ng et al., 2013; Ng et al., 2014). It captures word- and phrase-level edits by building an edit lattice and calculating an F-score over the lattice.

Felice and Briscoe (2015) note problems with M^2: specifically, it does not distinguish between a "do-nothing baseline" and systems that only propose wrong corrections; also, phrase-level edits can be easily gamed because the lattice treats the deletion of a long phrase as a single edit. To address these issues, they propose I-measure, which generates a token-level alignment between the source sentence, system output, and gold-standard sentences, and then computes accuracy based on the alignment.

Unlike these approaches, GLEU does not use error-coded references[4] (Napoles et al., 2015). Based on BLEU (Papineni et al., 2002), it computes n-gram precision of the system output against reference sentences. GLEU additionally penalizes text in the output that was unchanged from the source but changed in the reference sentences.

Recent work by Napoles et al. (2015) and Grundkiewicz et al. (2015) evaluated these metrics against human evaluations obtained using methods borrowed from the Workshop on Statistical Machine Translation (Bojar et al., 2014). Both papers found a moderate to strong correlation with human judgments for GLEU and M^2, and a slightly negative correlation for I-measure. Importantly, however, none of these metrics achieved as a high correlation with the human oracle ranking as desired in a fully reliable metric.

In Section 4, we examine the available metrics over different types of reference sets to identify an evaluation setup nearly as reliable as human experts.

3 Creating a new, *fluent* GEC corpus

We hypothesize that human judges, when presented with two versions of a sentence, will favor *fluent* versions over ones that exhibit only *technical grammaticality*.

By *technical grammaticality*, we mean adherence to an accepted set of grammatical conventions. In contrast, we consider a text to be *fluent* when it looks and sounds natural to a native-speaking population. Both of these terms are hard to define precisely, and fluency especially is a nuanced concept for which there is no checklist of criteria to be met.[5] To carry the intuitions, Table 1 contains examples of sentences that are one, both, or neither. A text does not have to be technically grammatical to be considered fluent, although in almost all cases, fluent texts are also technically grammatical. In the rest of this paper, we will demonstrate how they are quantifiably different with respect to GEC.

Annotating coded errors encourages a minimal set of edits because more substantial edits often address overlapping and interacting errors. For example, the annotators of the NUCLE corpus, which was used for the recent shared tasks, were explicitly instructed to select the minimal text span of possible alternatives (Dahlmeier et al., 2013). There are situations where error-coded annotations are useful to help students correct specific grammatical errors. The ability to do this with the non-error-coded, fluent annotations we advocate here is no longer direct, but is not lost entirely. For this purpose, some recent studies have proposed *post hoc* automated error-type classification methods (Swanson and Yamangil, 2012; Xue and Hwa, 2014), which compare the original sentence to its correction and deduce the error types.

We speculate that, by removing the error-coding restraint, we can obtain edits that sound more fluent to native speakers while also reducing the expense of annotation, with diminished time and training requirements. Chodorow et al. (2012) and Tetreault et al. (2014) suggested that it is better to have a large number of annotators to reduce bias in automatic evaluation. Following this recommendation, we collected additional annotations without error codes, written by both experts and non-experts.

[4] We use the term *references* to refer to the corrected sentences, since the term *gold standard* suggests that there is just one right correction.

[5] It is important to note that both grammaticality and fluency are determined with respect to a particular speaker population and setting. In this paper, we focus on Standard Written English, which is the standard used in education, business, and journalism. While judgments of individual sentences would differ for other populations and settings (for example, spoken African-American Vernacular English), the distinction between grammaticality and fluency would remain.

	Technically grammatical	Not technically grammatical
Fluent	In addition, it is impractical to make such a law.	I don't like this book, it's really boring.
Not fluent	Firstly , someone having any kind of disease belongs to his or her privacy .	It is unfair to release a law only point to the genetic disorder.

Table 1: Examples and counterexamples of *technically grammatical* and *fluent* sentences.

Original	Genetic disorder may or may not be hirataged hereditary disease and it is sometimes hard to find out one has these kinds of diseases .
Expert fluency	A genetic disorder may or may not be □ a hereditary disease , and it is sometimes hard to find out whether one has these kinds of diseases .
Non-expert fluency	Genetic □ factors can manifest overtly as disease □ , or simply be carried , making it □ hard , sometimes , to find out if one has □ a genetic predisposition to disease .

Table 2: An example sentence with expert and non-expert fluency edits. Moved and changed or inserted spans are underlined and □ indicates deletions.

3.1 Data collection

We collected a large set of additional human corrections to the NUCLE 3.2 test set, [6] which was used in the 2014 CoNLL Shared Task on GEC (Ng et al., 2014) and contains 1,312 sentences error-coded by two trained annotators. Bryant and Ng (2015) collected an additional eight annotations using the same error-coding framework, referred to here as BN15.

We collected annotations from both experts and non-experts. The experts[7] were three native English speakers familiar with the task. To ensure that the edits were clean and meaning-preserving, each expert's corrections were inspected by a different expert in a second pass. For non-experts, we used crowdsourcing, which has shown potential for annotating closed-class errors as effectively as experts (Tetreault et al., 2010; Madnani et al., 2011; Tetreault et al., 2014). We hired 14 participants on Amazon Mechanical Turk (MTurk) who had a HIT approval rate of at least 95% and were located in the United States. The non-experts went through an additional screening process: before completing the task, they wrote corrections for five sample sentences, which were checked by the three experts.[8]

We collected four complete sets of annotations by both types of annotators: two sets of *minimal edits*, designed to make the original sentences *technically grammatical* (following the NUCLE annotation instructions but without error coding), and two sets of *fluency edits*, designed to elicit native-sounding, *fluent* text. The instructions were:

- *Minimal edits:* Make the smallest number of changes so that each sentence is grammatical.
- *Fluency edits:* Make whatever changes necessary for sentences to appear as if they had been written by a native speaker.

In total, we collected 8 ($2 \times 2 \times 2$) annotations from each original sentence (minimal and fluency, expert and non-expert, two corrections each). Of the original 1,312 sentences, the experts flagged 34 sentences that needed to be merged together, so we skipped these sentences in our analysis and experiments. In the next two subsections we compare the changes made under both the fluency and minimal edit conditions (Section 3.2) and show how humans rate corrections made by experts and non experts in both settings (Section 3.3).

3.2 Edit analysis

When people (both experts and non-experts) are asked to make minimal edits, they make few changes

[6] www.comp.nus.edu.sg/~nlp/conll14st.html

[7] All of the expert annotators are authors of this work.

[8] The experts verified that the participants were following the instructions and not gaming the HITs.

Original	Some family may feel hurt , with regards to their family pride or reputation , on having the knowledge of such genetic disorder running in their family .
NUCLE	Some family members may feel hurt [] with regards to their family pride or reputation [] on having the knowledge of a genetic disorder running in their family .
Expert fluency	On [] learning of such a genetic disorder running in their family , some family members may feel hurt [] regarding their family pride or reputation .
Non-expert fluency	Some relatives may [] be concerned about the family 's [] reputation – not to mention their own pride – in relation to this news of [] familial genetic defectiveness [] .
Expert minimal	Some families may feel hurt [] with regards to their family pride or reputation , on having [] knowledge of such a genetic disorder running in their family .
Non-expert minimal	Some family may feel hurt [] with regards to their family pride or reputation [] on having the knowledge of such genetic disorder running in their family .

Table 3: An example sentence with the original NUCLE correction and fluency and minimal edits written by experts and non-experts. Moved and changed or inserted spans are underlined and [] indicates deletions.

to the sentences and also change fewer of the sentences. Fluency edits show the opposite effect, with non-experts taking more liberties than experts with both the number of sentences changed and the degree of change within each sentence (see Table 2 for an extreme example of this phenomenon).

In order to quantify the extent of changes made in the different annotations, we look at the percent of sentences that were left unchanged as well as the number of changes needed to transform the original sentence into the corrected annotation. To calculate the number of changes, we used a modified Translation Edit Rate (TER), which measures the number of edits needed to transform one sentence into another (Snover et al., 2006). An edit can be an insertion, deletion, substitution, or shift. We chose this metric because it counts the movement of a phrase (a *shift*) as one change, which the Levenshtein distance would heavily penalize. TER is calculated as the number of changes per token, but instead we report the number of changes per *sentence* for ease of interpretation, which we call the sTER.

We compare the original set of sentences to the new annotations and the existing NUCLE and BN15 reference sets to determine the relative extent of changes made by the fluency and minimal edits (Figure 2). Compared to the original, non-experts had a higher average sTER than experts, meaning that they made more changes per sentence. For fluency edits, experts and non-experts changed approximately the same number of sentences, but the non-experts made about seven edits per sentence compared to the experts' four. Minimal edits by both experts and non-experts exhibit a similar degree of change from the original sentences, so further qualitative assessment is necessary to understand whether the annotators differ. Table 3 contains an example of how the same ungrammatical sentence was corrected using both minimal and fluency edits, as well as one of the original NUCLE corrections.

The error-coded annotations of NUCLE and BN15 fall somewhere in between the fluency and minimal edits in terms of sTER. The most conservative set of sentences is the system output of the CoNLL 2014 shared task, with sTER $= 1.4$, or approximately one change made per sentence. In contrast, the most conservative human annotations, the minimal edits, edited a similar percent of the sentences but made about two changes per sentence.

When there are multiple annotators working on the same data, one natural question is the inter-annotator agreement (IAA). For GEC, IAA is often low and arguably not an appropriate measure of agreement (Bryant and Ng, 2015). Additionally, it would be difficult, if possible, to reliably calculate IAA without coded alignments between the new and

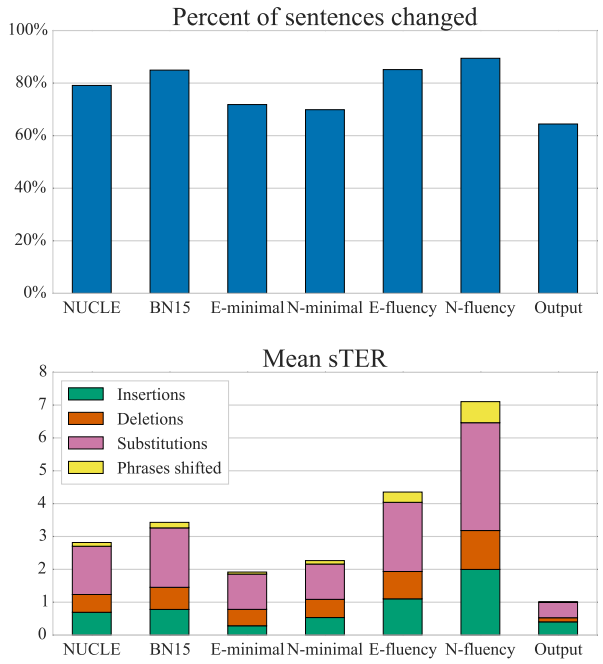

Figure 2: Amount of changes made by different annotation sets compared to the original sentences.

original sentences. Therefore, we look at two alternate measures: the percent of sentences to which different annotators made the same correction(s) and the sTER between two annotators' corrections, reported in Table 4.

As we expect, there is notably lower agreement between the annotators for fluency edits than for minimal edits, due to the presumably smaller set of required versus optional stylistic changes. Expert annotators produced the same correction on 15% of the fluency edits, but more than 38% of their minimal edits were identical. Half of these identical sentences were unchanged from the original. There was lower agreement between non-expert annotators than experts on both types of edits. We performed the same calculations between the two NUCLE annotators and found that they had agreement rates similar to the non-expert minimal edits. However, the experts' minimal edits have much higher consensus than both the non-experts' and NUCLE, with twice as many identical corrected sentences and half the sTER.

From this analysis, one could infer that the expert annotations are more reliable than the non-expert because there are fewer differences between annotators and fewer changes per sentence.

Edit type	Annotator	Identical	sTER
Fluency	E_1 v. E_2	15.3%	5.1
	N_1 v. N_2	5.9%	10.0
	E v. N	8.5%	7.9
Minimal	E_1 v. E_2	38.7%	1.7
	N_1 v. N_2	21.8%	2.9
	E v. N	25.9%	2.4
NUCLE	A v. B.	18.8%	3.3

Table 4: A comparison of annotations across different annotators (E for expert, N for non-expert). Where there were more than two annotators, statistics are over the full pairwise set. *Identical* refers to the percentage of sentences where both annotators made the same correction and *sTER* is the mean sTER between the annotators' corrections.

3.3 Human evaluation

As an additional validation, we ran a task to establish the relative quality of the new fluency and minimal-edit annotations using crowdsourcing via MTurk. Participants needed to be in the United States with a HIT approval rate of at least 95% and pass a preliminary ranking task, graded by the authors. We randomly selected 300 sentences and asked participants to rank the new annotations, one randomly selected NUCLE correction, and the original sentence in order of grammaticality and meaning preservation (that is, a sentence that is well-formed but changes the meaning of the original source should have a lower rank than one that is equally well-formed but maintains the original meaning). Since we were comparing the minimal edits to the fluency edits, we did not define the term *grammaticality*, but instead relied on the participants' understanding of the term. Each sentence was ranked by two different judges, for a total of 600 rankings, yielding 7,795 pairwise comparisons.

To rank systems, we use the TrueSkill approach (Herbrich et al., 2006; Sakaguchi et al., 2014), based on a protocol established by the Workshop on Machine Translation (Bojar et al., 2014; Bojar et al., 2015). For each competing system, TrueSkill infers the absolute system quality from the pairwise comparisons, representing each as the mean of a Gaussian. These means can then be sorted to rank sys-

#	Score	Range	Annotation type
1	1.164	1–2	Expert fluency
	0.976	1–2	Non-expert fluency
3	0.540	3	NUCLE
4	0.265	4	Expert minimal
5	-0.020	5	Non-expert minimal
6	-2.925	6	Original sentence

Table 5: Human ranking of the new annotations by grammaticality. Lines between systems indicate clusters according to bootstrap resampling at $p \leq 0.05$. Systems in the same cluster are considered to be tied.

tems. By running TrueSkill 1,000 times using bootstrap resampling and producing a system ranking each time, we collect a range of ranks for each system. We can then cluster systems according to non-overlapping rank ranges (Koehn, 2012) to produce the final ranking, allowing ties.

Table 5 shows the ranking of "grammatical" judgments for the additional annotations and the original NUCLE annotations. While the score of the expert fluency edits is higher than the non-expert fluency, they are within the same cluster, suggesting that the judges perceived them to be just as good. The fluency rewrites by both experts and non-experts are clearly preferable over the minimal edit corrections, although the error-coded NUCLE corrections are perceived as more grammatical than the minimal corrections.

4 Automatic metrics

We have demonstrated that humans prefer fluency edits to error-coded and minimal-edit corrections, but it is unclear whether these annotations are an effective reference for automatic evaluation. The broad range of changes that can be made with non-minimal edits may make it especially challenging for current automatic evaluation metrics to use. In this section, we investigate the impact that different reference sets have on the system ranking found by different evaluation metrics. With reference sets having such different characteristics, the natural question is: which reference and evaluation metric pairing best reflects human judgments of grammaticality?

To answer this question, we performed a comprehensive investigation of existing metrics and annotation sets to evaluate the 12 system outputs made public from the 2014 CoNLL Shared Task. To our knowledge, this is the first time that the interplay of annotation scheme and evaluation metric, as well as the rater expertise, has been evaluated jointly for GEC.

4.1 Experimental setup

The four automatic metrics that we investigate are M^2, I-measure,[9] GLEU, and BLEU. We include the machine-translation metric BLEU because evaluating against our new non-coded annotations is similar to machine-translation evaluation, which considers overlap instead of absolute alignment between the output and reference sentences.

For the M^2 and I-measure evaluations, we aligned the fluency and minimal edits to the original sentences using a Levenshtein edit distance algorithm.[10] Neither metric makes use of the annotation labels, so we simply assigned dummy error codes.

Our GLEU implementation differs from that of Napoles et al. (2015). We use a simpler, modified version: Precision is the number of candidate (C) n-grams that match the reference (R) n-grams, minus the counts of n-grams found more often in the source (S) than the reference (Equation 1). Because the number of possible reference n-grams increases as more reference sets are used, we calculate an intermediate GLEU by drawing a random sample from one of the references and report the mean score over 500 iterations.[11]

We compare the system outputs to each of the six annotation sets and a seventh set containing all of the annotations, using each metric. We ranked the systems based on their scores using each metric–annotation-set pair, and thus generated a total of 28 different rankings (4 metrics $\times$ 7 annotation sets).

To determine the best metric, we compared the system-level ranking obtained from each evaluation technique against the expert human ranking reported in Grundkiewicz et al. (2015), Table 3c.

[9] We ran I-measure with the `-nomix` flag, preventing the algorithm from finding the optimal alignment across all possible edits. Alignment was very memory-intensive and time consuming, even when skipping long sentences.

[10] Costs for insertion, deletion, and substitution are set to 1, allowing partial match (e.g. same lemma).

[11] Running all iterations, it takes less than 30 seconds to evaluate 1,000 sentences.

$$p_n^* = \frac{\left(\sum_{ngram \in \{C \cap R\}} count_{C,R}(ngram) - \sum_{ngram \in \{C \cap S\}} \max\left[0, count_{C,S}(ngram) - count_{C,R}(ngram)\right] \right)}{\sum_{ngram \in \{C\}} count(ngram)}$$

$$count_{A,B}(ngram) = \min\left(\#\text{ occurrences of } ngram \text{ in } A, \#\text{ occurrences of } ngram \text{ in } B\right) \quad (1)$$

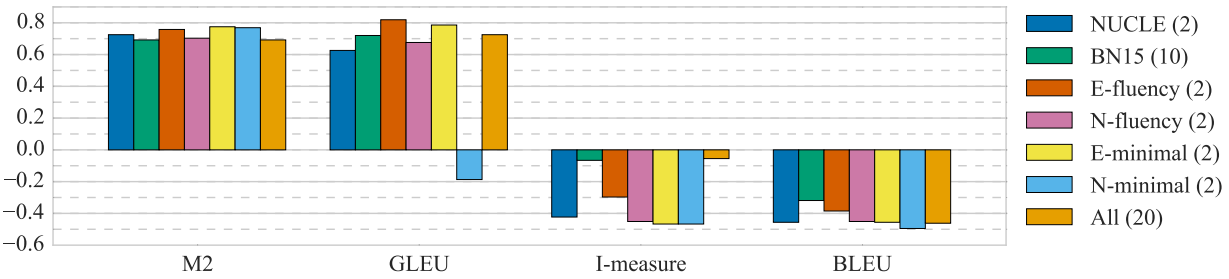

Figure 3: Correlation of the human ranking with metric scores over different reference sets (Spearman's ρ). The number of annotations per sentence in each set is in parentheses. See Table 6 for the numeric values.

	M^2	GLEU	I-measure	BLEU
NUCLE	0.725 0.677*	0.626 0.646	-0.423 -0.313	-0.456 -0.310
BN15	0.692 0.641	0.720 0.697	-0.066 -0.007	-0.319* -0.255
E-fluency	0.758 0.665	**0.819*** **0.731***	-0.297 -0.256	-0.385 -0.230*
N-fluency	0.703 0.655	0.676 0.668	-0.451 -0.319	-0.451 -0.388
E-min.	0.775* 0.655	0.786 0.676	-0.467 -0.385	-0.456 -0.396
N-min.	0.769 0.641	-0.187 -0.110	-0.467 -0.402	-0.495 -0.473
All	0.692 0.617	0.725 0.724	-0.055* 0.061*	-0.462 -0.314

Table 6: Correlation between the human ranking and metric scores over different reference sets. The first line of each cell is Spearman's ρ and the second line is Pearson's r. The strongest correlations for each metric are starred, and the overall strongest correlations are in bold.

4.2 Results

Figure 3 and Table 6 show the correlation of the expert rankings with all of the evaluation configurations. For the leading metrics, M^2 and GLEU, the expert annotations had stronger positive correlations than the non-expert annotations. Using just two expert fluency annotations with GLEU has the strongest correlation with the human ranking out of all other metric–reference pairings ($\rho = 0.819$, $r = 0.731$), and it is additionally cheaper and faster to collect. E-fluency is the third-best reference set with M^2, which does better with minimal changes: the reference sets with the strongest correlations for M^2 are E-minimal ($\rho = 0.775$) and NUCLE ($r = 0.677$). Even though the non-expert fluency edits had more changes than the expert fluency edits, they still did reasonably well using both M^2 and GLEU.

The GLEU metric has strongest correlation when comparing against the E-fluency, BN15, E-minimal, and "All" reference sets. One could argue that, except for E-minimal, these references all have greater diversity of edits than NUCLE and minimal edits. Although BN15 has fewer changes made per sentence than the fluency edits, because of the number of annotators, the total pool of n-grams seen per sentence increases. E-minimal edits also have strong correlation, suggesting there may be a trade-off between quantity and quality of references.

A larger number of references could improve performance for GLEU. Because fluency edits tend to have more variations than error-coded minimal-edit annotations, it is not obvious how many fluency edits are necessary to cover the full range of possible corrections. To address this question, we ran an ad-

ditional small-scale experiment, where we collected 10 non-expert fluency edits for 20 sentences and computed the average GLEU scores of the submitted systems against an increasing number of these fluency references. The result (Figure 5) shows that the GLEU score with more fluency references, but the effect starts to level off when there are at least 4 references, suggesting that 4 references cover the majority of possible changes. A similar pattern was observed by Bryant and Ng (2015) in error-coded annotations with the M^2 metric.

The reference sets against which M^2 has the strongest correlation are NUCLE, expert fluency, and expert minimal edits. Even non-expert fluency annotations result in a stronger correlation with the human metric than BN15. These findings support the use of fluency edits even with a metric designed for error-coded corpora.

One notable difference between M^2 and GLEU is their relative performance using non-expert minimal edits as a metric. M^2 is robust to the non-expert minimal edits and, as a reference set, this achieves the second strongest Spearman's correlation for this metric. However, pairing the non-expert minimal edits with GLEU results in slightly *negative* correlation. This is an unexpected result, as there is sizable overlap between the non-expert and expert minimal edits (Table 4). We speculate that this difference may be due to the quality of the non-expert minimal edits. Recall that humans perceived these sentences to be worse than the other annotations, and better only than the original sentence (Table 5).

I-measure and BLEU are shown to be unfavorable for this task, having negative correlation with the human ranking, which supports the findings of Napoles et al. (2015) and Grundkiewicz et al. (2015). Even though BLEU and GLEU are both based on the n-gram overlap between the hypothesis and original sentences, GLEU has strong positive correlations with human rankings while BLEU has a moderate negative correlation. The advantage of GLEU is that it penalizes n-grams in the system output that were present in the input sentence and absent from the reference. In other words, a system loses credit for missing n-grams that should have been changed. BLEU has no such penalty and instead only rewards n-grams that occur in the references and the output, which is a problem in same-language text rewriting tasks where there is significant overlap between the reference and the original sentences. For this data, BLEU assigns a higher score to the original sentences than to any of the systems.[12]

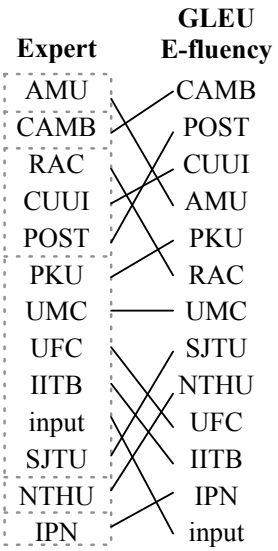

Figure 4: System rankings produced by GLEU with expert fluency (E-fluency) as the reference compared to the expert human ranking.

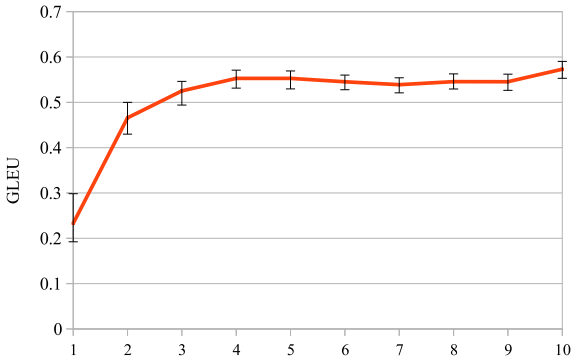

Figure 5: Mean GLEU scores with different numbers of fluency references. The red line corresponds to the average GLEU score of the 12 GEC systems and the vertical bars show the maximum and minimum GLEU scores.

Figure 4 shows the system ranking for the most strongly correlated annotation–evaluation combination (GLEU with E-fluency) compared to the "ground truth" human rankings. The automatic metric clusters the systems into the correct upper and lower halves, and the input is correctly placed in the lower half of the rankings.

[12]Of course, it could be that the input sentences are the best, but the human ranking in Figure 4 suggests otherwise.

Even though automatic metrics strongly correlate with human judgments, they still do not have the same reliability as manual evaluation. Like error-coded annotations, judgment by specialists is expensive, so we investigate a more practical alternative in the following section.

5 Human evaluation

Automatic metrics are only a proxy for human judgments, which are crucial to truthfully ascertain the quality of systems. Even the best result in Section 4.2, which is state of the art and has very strong rank correlation ($\rho = 0.819$) with the expert ranking, makes dramatic errors in the system ranking. Given the inherent imperfection of automatic evaluation (and possible over-optimization to the NUCLE data set), we recommend that human evaluation be produced alongside metric scores whenever possible. However, human judgments can be expensive to obtain. Crowdsourcing may address this problem and has been shown to yield reasonably good judgments for several error types at a relatively low cost (Tetreault et al., 2014). Therefore, we apply crowdsourcing to sentence-level grammaticality judgments, by replicating previous experiments that reported expert rankings of system output (Napoles et al., 2015; Grundkiewicz et al., 2015) using non-experts on MTurk.

5.1 Experimental setup

Using the same data set as those experiments and the work described in this paper, we asked screened participants[13] on MTurk to rank five randomly selected systems and NUCLE corrections from best to worst, with ties allowed. 294 sentences were randomly selected for evaluation from the NUCLE subsection used in Grundkiewicz et al. (2015), and the output for each sentence was ranked by two different participants. The 588 system rankings yield 26,265 pairwise judgments, from which we inferred the absolute system ranking using TrueSkill.

5.2 Results

Figure 6 compares the system ranking by non-experts to the same expert ranking used in Section 4.1. The rankings have very strong correlation ($\rho = 0.917$, $r = 0.876$), indicating that non-expert grammaticality judgments are comparably as reliable as those by experts. Compared to the best metric ranking shown in Figure 4, the non-expert ranking appears significantly better. No system has a rank more than two away from the expert rank, while GLEU has six systems with ranks that are three away. The non-expert correlation can be seen as an upper bound for the task, which is approached but not yet attained by automatic metrics.

Systems in the same cluster, indicated by dotted lines in Figure 6, can be viewed as ties. From this perspective the expert and non-expert rankings are virtually identical. In addition, experts and non-experts have similar inter-annotator agreement in their pairwise system judgments (Table 7). The agreement between experts and non-experts is lower than the agreement between just experts or just non-

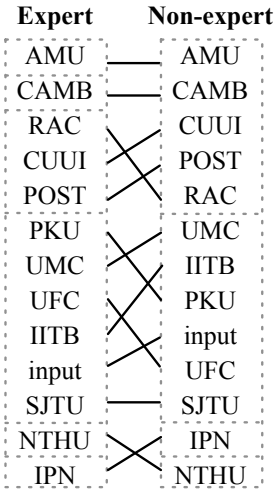

Figure 6: Output of system rankings by experts and non-experts, from best to worst. Dotted lines indicate clusters according to bootstrap resampling ($p \leq 0.05$).

Judges	κ	κ_w
Non-experts	0.29	0.43
Experts	0.29	0.45
Non-experts and Experts	0.15	0.23

Table 7: Inter-annotator agreement of pairwise system judgments within non-experts, experts and between them. We show Cohen's κ and quadratic-weighted κ.[15]

[13]Participants in the United States with a HIT approval rate $\geq 95\%$ had to pass a sample ranking task graded by the authors.

[15]In addition to Cohen's κ, we report weighted κ because $A > B$ and $A < B$ should have less agreement than $A > B$ and $A = B$.

experts, which may be due to the difference of these experimental settings for experts (Grundkiewicz et al., 2015) and for non-experts (this work). However, this finding is not overly concerning since the correlation between the rankings is so strong.

In all, judgments cost approximately $140 ($0.2 per sentence) and took a total of 32 hours to complete. Because the non-expert ranking very strongly correlates to the expert ranking and non-experts have similar IAA as experts, we conclude that expensive expert judgments can be replaced by non-experts, when those annotators have been appropriately screened.

6 Conclusion

There is a real distinction between technical grammaticality and fluency. Fluency is a level of mastery that goes beyond knowledge of how to follow the rules, and includes knowing when they can be broken or flouted. Language learners—who are a prime constituency motivating the GEC task—ultimately care about the latter. But crucially, the current approach of collecting error-coded annotations places downward pressure on annotators to minimize edits in order to neatly label them. This results in annotations that are less fluent, and therefore less useful, than they should be. We have demonstrated this with the collection of both minimally-edited and fluent rewrites of a common test set (Section 3.1); the preference for fluent rewrites over minimal edits is clear (Table 5).

To correct this, the annotations and associated metrics used to score automated GEC systems should be brought more in line with this broadened goal. We advocate for the collection of fluent sentence-level rewrites of ungrammatical sentences, which is cheaper than error-coded annotations and provides annotators with the freedom to produce fluent edits. In the realm of automatic metrics, we found that a modified form of GLEU computed against expert fluency rewrites correlates best with a human ranking of the systems; a close runner-up collects the rewrites from non-experts instead of experts.

Finally, to stimulate metric development, we found that we were able to produce a new human ranking of systems using non-expert judges. These judges produced a ranking that was highly correlated with the expert ranking produced in earlier work (Grundkiewicz et al., 2015). The implication is further reduced costs in producing a gold-standard ranking for new sets of system outputs against both existing and new corpora.

As a result, we make the following recommendations:

- GEC should be evaluated against 2–4 whole-sentence rewrites, which can be obtained by non-experts.
- Automatic metrics that rely on error coding are not necessary, depending on the use case. Of the automatic metrics that have been proposed, we found that a modified form of GLEU (Napoles et al., 2015) is the best-correlated.
- The field of GEC is in danger from over-reliance on a single annotated corpus (NUCLE). New corpora should be produced in a regular fashion, similar to the Workshop on Statistical Machine Translation.

Fortunately, collecting annotations in the form of unannotated sentence-level rewrites is much cheaper than error-coding, facilitating these practices.

By framing grammatical error correction as fluency, we can reduce the cost of annotation while creating a more reliable gold standard. We have clearly laid improved practices for annotation and evaluation, demonstrating that better quality results can be achieved for less cost using fluency edits instead of error coding. All of the source code and data, including templates for data collection, will be publicly available, which we believe is crucial for supporting the improvement of GEC in the long term.

Acknowledgments

We would like to thank Christopher Bryant, Mariano Felice, Roman Grundkiewicz and Marcin Junczys-Dowmunt for providing data and code. We would also like to thank the TACL editor, Chris Quirk, and the three anonymous reviewers for their comments and feedback. This material is based upon work partially supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1232825.

References

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.

Christopher Bryant and Hwee Tou Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 697–707, Beijing, China, July. Association for Computational Linguistics.

Martin Chodorow and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, pages 140–147.

Martin Chodorow, Markus Dickinson, Ross Israel, and Joel Tetreault. 2012. Problems in evaluating grammatical error detection systems. In *Proceedings of COLING 2012*, pages 611–628, Mumbai, India, December. The COLING 2012 Organizing Committee.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada, June. Association for Computational Linguistics.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner english: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia, June. Association for Computational Linguistics.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 242–249, Nancy, France, September. Association for Computational Linguistics.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62, Montréal, Canada, June. Association for Computational Linguistics.

Mariano Felice and Ted Briscoe. 2015. Towards a standard evaluation method for grammatical error detection and correction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, Denver, CO, June. Association for Computational Linguistics.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal, September. Association for Computational Linguistics.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill™: A Bayesian Skill Rating System. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, pages 569–576, Vancouver, British Columbia, Canada, December. MIT Press.

Philipp Koehn. 2012. Simulating human judgment in machine translation evaluation campaigns. *Proceedings IWSLT 2012*.

C. Leacock, M. Chodorow, M. Gamon, and J. Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners, Second Edition*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Nitin Madnani, Martin Chodorow, Joel Tetreault, and Alla Rozovskaya. 2011. They can help: Using crowdsourcing to improve the evaluation of grammatical error detection systems. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 508–513, Portland, Oregon, USA, June. Association for Computational Linguistics.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. Gleu: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic, June. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical

error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China, July. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria, August. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland, June. Association for Computational Linguistics.

Diane Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference*, pages 572–581.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Ellie Pavlick, Rui Yan, and Chris Callison-Burch. 2014. Crowdsourcing for grammatical error correction. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 209–212. ACM.

Alla Rozovskaya and Dan Roth. 2010. Annotating ESL errors: Challenges and rewards. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36, Los Angeles, California, June. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2014. Building a state-of-the-art grammatical error correction system. *Transactions of the Association for Computational Linguistics*, 2:414–434.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Ben Swanson and Elif Yamangil. 2012. Correction detection and error type selection as an ESL educational aid. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 357–361, Montréal, Canada, June. Association for Computational Linguistics.

Joel Tetreault and Martin Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 24–32, Manchester, UK, August. Coling 2008 Organizing Committee.

Joel Tetreault, Elena Filatova, and Martin Chodorow. 2010. Rethinking grammatical error annotation and evaluation with the Amazon Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–48, Los Angeles, California, June. Association for Computational Linguistics.

Joel Tetreault, Martin Chodorow, and Nitin Madnani. 2014. Bucking the trend: Improved evaluation and annotation practices for ESL error detection systems. *Language Resources and Evaluation*, 48(1):5–31.

Huichao Xue and Rebecca Hwa. 2014. Improved correction detection in revised ESL sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–604, Baltimore, Maryland, June. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA, June. Association for Computational Linguistics.